

Automatic Stress Detection in Working Environments from Smartphones' Accelerometer Data: A First Step

Enrique Garcia-Ceja, Venet Osmani and Oscar Mayora

Abstract—Increase in workload across many organisations and consequent increase in occupational stress is negatively affecting the health of the workforce. Measuring stress and other human psychological dynamics is difficult due to subjective nature of self-reporting and variability between and within individuals. With the advent of smartphones it is now possible to monitor diverse aspects of human behaviour, including objectively measured behaviour related to psychological state and consequently stress. We have used data from the smartphone's built-in accelerometer to detect behaviour that correlates with subjects stress levels. Accelerometer sensor was chosen because it raises fewer privacy concerns (in comparison to location, video or audio recording, for example) and because its low power consumption makes it suitable to be embedded in smaller wearable devices, such as fitness trackers. 30 subjects from two different organizations were provided with smartphones. The study lasted for 8 weeks and was conducted in real working environments, with no constraints whatsoever placed upon smartphone usage. The subjects reported their perceived stress levels three times during their working hours. Using combination of statistical models to classify self reported stress levels, we achieved a maximum overall accuracy of 71% for user-specific models and an accuracy of 60% for the use of similar-users models, relying solely on data from a single accelerometer.

Index Terms—automatic stress detection, health monitoring, accelerometer, smartphones, ambient intelligence, health and well-being.

I. INTRODUCTION

THE competitive nature of the world economy and the use of advanced information and communication technologies has changed the nature of workplace environments, ensuring increased connectivity and consequently reachability of workers even outside working hours. This has resulted in an increase of workload [1], which has become a common issue in many organisations, where employees experience psychological problems related to occupational stress. According to the Fourth European Working Conditions Survey (EWCS), work-related stress was reported by 22% of workers from 27 Member states of the European Union [2]. Furthermore, higher

E. Garcia-Ceja is a PhD student at Tecnológico de Monterrey, Monterrey, México. e-mail: e.g.mx@ieee.org

V. Osmani and O. Mayora are with CREATE-NET International Research Centre, Trento, Italy. e-mail: venet.osmani@create-net.org oscar.mayora@create-net.org

Author's copy of the paper published in IEEE Journal of Biomedical and Health Informatics

Enrique Garica-Ceja, Venet Osmani, Oscar Mayora "Automatic Stress Detection in Working Environments from Smartphones Accelerometer Data: A First Step" IEEE Journal of Biomedical and Health Informatics, DOI:10.1109/JBHI.2015.2446195, 2015

prevalence of stress has been reported in North America, where 55% of population has reported increased workload having a significant impact on physical and mental health as described in APA Survey [3].

Occupational stress has been proven to contribute to disease activation. Several research studies have found that stress at work is associated with cardiovascular diseases [4], musculoskeletal diseases [5], immunological problems [6], and problems with mental health such as anxiety and depression disorders [7]. In regard to organizational well-being, a decline of physical and mental health of workers has been reported in Paoli et al.[8], leading to a decrease in the performance, decrease in overall productivity of organization and increased cost in terms of absenteeism. Experiencing work-related stress is common in working environments and low levels of stress can even result in productivity increase [9]. However, stress responses of employees are triggered when work-related pressure (such as quantity of work to be accomplished in a short period of time, pressure to work overtime, low social support, job insecurity and lesser breaks or holidays) challenge the human ability to cope with them.

Considering detrimental effects of prolonged exposure to stress both for employees and organizations, there is a clear need for a system that can continuously monitor behaviour of workers and correlate various behaviour aspects with perceived stress levels. Several research works have used different sensing technologies, such as sound analysis [10], image processing from cameras [11] and physiological sensors [12] to detect stress. Considering privacy concerns when using cameras and microphones, physiological measures have become an increasingly popular approach for measuring stress-related signs from sensor data (typically GSR and heart-rate sensors), such as work in [13]. However, there are several concerns about using physiological sensors, principally due to their obtrusive nature, lack of comfort and ability to be worn continuously [14], consequently impacting natural behaviour of the subjects.

With these points in mind, and based on our previous studies [15], [16], smartphones have a distinct advantage in that they are already familiar and widely adopted devices, thus minimising "observer effect" and do not pose additional discomfort on the monitored subjects [17], [18]. Using smartphones to monitor behaviour of subjects, we report the results of our study in detecting stress levels in real working environments. We recruited 30 subjects from two different organizations that participated in our 8 week study, where each participant

reported perceived stress levels three times during working hours using self-assessment questionnaire.

Through the use of a combination of statistical models to classify self reported stress levels, we achieved an overall accuracy of 71% for *user-specific* models and 60% for the use of *similar-users* models. These results are comparable to the state of the art results in stress recognition, with the difference that our work relies solely on a single triaxial accelerometer sensor. Furthermore, we also developed classification models using data from similar users, when building individual models was not feasible due to scarce data. Lastly, we evaluate the use of an ordinal classifier to take into account the class ordering information of the different stress levels.

Relying on a single accelerometer as the only sensor in detecting stress is especially promising when considering exponential rise of personal activity trackers (such as FitBit or Jawbone) that typically contain a single embedded accelerometer.

The rest of the paper is organized as follows: Section II summarises previous research works for monitoring stress events from individuals in work- and real-life settings. Section III provides information about the group selection for the study, and how the data was collected. The details of data preprocessing are given in Section IV. Section V presents an exploratory data analysis as a first step towards building statistical classification models. Section VI presents the details of the different schemes used to classify stress levels. Section VII describes the experiments and results of our study. Conclusions and future research directions are given in Section VIII.

II. RELATED WORK

There have been several works that aim to detect stress in an automatic manner. For example, Carneiro et al. [19] used video cameras, accelerometers, touchscreens to extract different features while inducing different levels of stress during an electronic game session. Their experiments included 19 subjects and they used a J48 tree to classify touches as stressed or not achieving an accuracy of 78%. In [20], Giakoumis et al. used video, accelerometers at the user's knees, galvanic skin response and electrocardiogram sensors to detect stress. There were 21 participants in their study and the Stroop color test [21] was used to induce stress. Their results showed that using behavioural features together with physiological measures helped to increase the stress detection accuracy compared when using just physiological features. Recently, there has also been research to detect stress outside lab environments by using wearable sensors. Lu et al. [22] implemented an application running in a smartphone to detect stress using voice as input. Sano & Picard [17] used data collected from a wrist sensor, surveys and a mobile phone to classify stressed and not stressed states achieving results of over 75% accuracy.

Two types of setups that have been used in previous works can be identified: *In-lab experiments* and *unconstrained experiments*. In-lab experiments are performed with controlled conditions, i.e., subjects are required to stay within a specific physical place and to follow a standard protocol. This protocol

generally consists of filling surveys and performing a series of experiments in a specific order. In an unconstrained setup, the subject is generally given a set of wearable sensors and the data is collected while the subject performs their daily routines without following any predefined schedule.

Table I presents a summary of related works on automatic stress detection and classified according to the type of experiment: *In-lab*, *unconstrained* and the type of stressors: *controlled*, *uncontrolled* and *unknown*. This work differs from the previous work in the following aspects: 1) The data was collected in an unconstrained out of the lab environment and with unknown stressors using only an accelerometer sensor from smartphone; 2) We explore the potential of using data from a single source (accelerometer) to detect acute stress levels. We chose this sensor because it is non-visual and non-auditory, and thus mitigates privacy concerns and does not interfere with the individual's daily routines [23], [24], [25]; and 3) We built classification models using data from similar users in cases when building individual models is not feasible due to scarce data.

III. DATA COLLECTION

Behavioural data were obtained using the built-in sensors of Samsung Galaxy SIII Mini smartphones. The data were collected with the written, informed consent of all participants and stored in the memory of the smartphone using the application developed by our team. Additional information pertaining to the usage of apps and contextual information such as location, accelerometer, social activities, phone calls, SMS, Wi-Fi, and proximity was also recorded. However, in this work we analysed only the data from the triaxial accelerometer, recorded continuously. Given that the phone application collected data from several sensors, the accelerometer sampling rate was set at 5 Hz in order to optimise the battery life. This was adequate for our analysis since work in [29] showed that with a sampling rate of 5 Hz it is still possible to recognize physical activities with an accuracy of 94.98%, while we are not analysing short, fine grained movements (as in activity or gesture recognition) but rather focus on the overall behaviour that spans several minutes.

We also collected subjective information related to subjects' stress and psychological states involving a series of questions/answers gathered from a survey. These questions were derived from a clinically validated burnout questionnaire; the Oldenburg Burnout Inventory (OLBI) [12]. Subjective psychological scores for stress, were reported in a questionnaire three times during the working days (morning, afternoon, and end of workday) on a 5-point scale. This information was then converted into an ordinal scale to represent stress levels as *low*, *medium* and *high*, due to inherent differences in subjective reporting of stress levels between individuals and also within individuals [30], that is, for one user the value of 4 may mean 'highly stressed' whereas for another 'a little bit above normal'. Grouping the ratings into a smaller number of ordinal points alleviates some of the inherent subjectivity.

TABLE I

CLASSIFICATION OF DIFFERENT RELATED WORKS. TYPE COLUMN INDICATES IF THE EXPERIMENT WAS PERFORMED IN-LAB OR IN AN UNCONSTRAINED ENVIRONMENT AND THE TYPE OF STRESSORS USED: CONTROLLED, UNCONTROLLED, UNKNOWN.

Work	Type	Data sources	Details
Carneiro et al. [19]	In-lab, controlled	video cameras, accelerometers, pressure-sensitive touchscreens	19 subjects. 78% accuracy in classifying touches as stressed or not using a J48 tree.
Giakoumis et al. [20]	In-lab, controlled	video, accelerometers at users' knees, Galvanic skin response, electrocardiogram	21 subjects. Avg. accuracy of 100% for their dataset 1 and 96.6% for dataset 2 when using all sensors.
Sun et al. [26]	In-lab, controlled	electrocardiogram, galvanic skin response, accelerometer	20 subjects. Overall accuracy 92.4% for 10-fold cross validation and 80.9% between subjects classification.
Bauer & Lukowicz [18]	unconstrained, uncontrolled	gps, wi-fi, bluetooth, call logs, sms	7 subjects. Detected a change of behaviour during stress periods of approx. 86% of the participants.
Lu et al. [22]	In-lab, unconstrained, uncontrolled	audio	14 subjects. accuracy of 81% and 76% for indoor and outdoor environments with model adaptation.
Muaremi et al. [27]	unconstrained, unknown	heart rate, audio, acceleration, gps, calls, contacts, etc.	35 subjects. 61% accuracy for user specific models and 53% for a general model.
Sano & Picard [17]	unconstrained, unknown	accelerometer, skin conductance, calls, sms, location, screen	18 subjects. Accuracies over 75%
Bogomolov et al. [28]	unconstrained, unknown	call logs, sms, bluetooth	117 subjects. An overall recognition accuracy of 72.39% with Random forest model.

A. Participants

Sensor data was collected from 30 healthy subjects and analysed with self-reported stress data for a period of 8 weeks, excluding weekends. Due to user compliance issues, the average number of data collection days per user was 29 ± 6 . Furthermore, some surveys during the day were occasionally skipped by the users. The participants used the phone from morning until the end of the work day, without any restrictions whatsoever placed upon the use of the phone in a specific manner. In order to get insights in the working style and gain more knowledge from employees in their working environments, we chose to recruit participants from two different companies located in province of Trentino, Italy. The study involved 18 (60%) males and 12 (40%) females aged 37.46 ± 7.26 years. Participants were informed that the goal of the study was to monitor behaviour activities relevant to stress. All participants

consented to participate in the study and to have their data recorded. They were also informed that all the collected data was anonymous and will be used for research purposes only.

IV. PRE-PROCESSING

Feature extraction

From the raw accelerometer data a total of 34 features from time and frequency domain were extracted. The feature extraction was performed on non-overlapping fixed length windows of 128 samples (25.6 seconds.). The 34 features were: *Mean x axis*, *Mean y axis*, *Mean z axis*, *StdDev x axis*, *StdDev y axis*, *StdDev z axis*, *Variance x axis*, *Variance y axis*, *Variance z axis*, *Variance 3 axes*, *Mean 3 axes*, *Max 3 axes*, *Min 3 axes*, *Standard Deviation 3 axes*, *Absolute Value 3 axes*, *Median 3 axes*, *Range 3 axes*, *Variance Sum* [31], *Magnitude* Eq.(1), *Signal Magnitude Area* Eq.(2), *Root Mean Squared* Eq.(3), *Curve Length* Eq.(4), *Non Linear Energy* [32], *Entropy*: differential entropy from time domain magnitude Eq.(5) [33], *Energy*: which is the sum of the squared discrete FFT component magnitudes of the signal. Eq.(6) [34], *Mean Energy*, *StdDev Energy*, *DFT* (*Discrete Fourier Transform*), *Peak Magnitude* which is the maximum value of the magnitude. Eq.(7), *Peak Magnitude Frequency* which is the frequency that corresponds to the maximum magnitude. Eq.(8), *Peak Power* which is analogous to peak magnitude but on the power spectrum, *Peak Power Frequency* this is analogous to peak magnitude frequency, *Magnitude Entropy* Eq.(9) and *Power Shannon Entropy* same as *Magnitude Entropy* but over the power spectrum.

$$\text{Magnitude} = \frac{1}{n} \sum_{i=1}^n \sqrt{x_i^2 + y_i^2 + z_i^2} \quad (1)$$

$$SMA = \frac{1}{T} \int_o^T |x(t)| dt + \int_o^T |y(t)| dt + \int_o^T |z(t)| dt \quad (2)$$

$$RMS = \sqrt{\frac{1}{n} (x_1^2 + x_2^2 + \dots + x_n^2)} \quad (3)$$

$$curvelength = \sum_{i=2}^N |x_{i-1} - x_i| \quad (4)$$

$$h(X) = \int_{\mathbb{X}} f(x) \log f(x) dx \quad (5)$$

$$energy = \sum_{i=1}^{(n/2)} x[i]^2 \quad (6)$$

$$pm = \max_{i=1..(n/2)} x_i \quad (7)$$

$$pmf = \arg \max_{i=1..(n/2)} x_i \quad (8)$$

$$H(X) = - \sum_{i=0}^{N-1} p_i \log_2 p_i \quad (9)$$

Self-reported Stress

The stress scale in the survey has the scale 1 to 5, where 1 means least stressed and 5 means most stressed. For the purpose of our analysis we grouped those values into three groups: *low stress* for values of 1 and 2; *medium stress* for a value of 3; and *high stress* for values of 4 and 5.

For our analysis, we considered observations from the second and third surveys only because there is no accelerometer data before the first survey (beginning of the day). To characterize each survey, we took the features from the previous 2 hours for each survey and computed summary statistics which will be used as the final features: mean, maximum and minimum value of each of the 34 features giving a total of 102 features. Table II shows the total number of observations for each of the stress levels.

TABLE II
TOTAL NUMBER OF OBSERVATIONS FOR THE SECOND AND THIRD SURVEYS

Stress level:	Low	Medium	High	
# observations	667	521	329	Total: 1,517

V. EXPLORATORY DATA ANALYSIS

In this section we present a general overview of the data. Figure 1 shows the average of the self reported stress level scores by weekday over all users using data from the 3 surveys. It can be seen that the maximum stress level is reported on Tuesday and then begins to decrease towards its minimum on Friday. The resulting standard error bars overlap with each other suggesting that the differences between days are not significant, confirmed with an analysis of variance test.

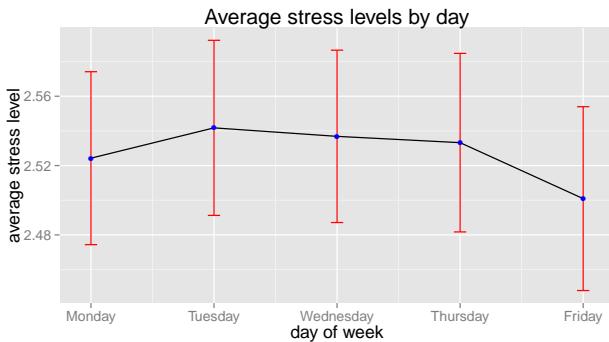


Fig. 1. Average stress levels by weekday with standard error bars of the mean.

Now we investigate whether extracted features could be used as potential predictors for stress levels. Figure 2 shows the estimated density function for the *Entropy* feature over all users. Vertical lines indicate the median. Through visual inspection, the difference between the median of the Entropy for high stress is clearly visible from that of low stress. The difference between medium and low stress is also clear and the difference between high and medium is still noticeable but smaller. It seems that Entropy is a good candidate feature (independently of the others) to differentiate between high/low

and medium/low stress levels but it may have difficulties differentiating between high/medium stress levels.

To see whether or not a specific feature is significantly different between every pair of possible stress levels (low/high, low/medium and high/medium) for each of the users, a Mann Whitney U test [35] was performed with a significance level $\alpha = 0.01$ and bonferroni p-values correction. This test was chosen because it is non-parametric and most of the feature distributions are not normal. The results of the statistical test indicated that for most of the features and users the differences were significant (except for the Peak Magnitude feature). However, this does not necessarily mean that most of the features will be good predictors since the differences may be too small to be detected or to be useful to a given classifier model. In order to check the effect size of each of the features we computed the Cohen's d effect size and quantified it using the thresholds defined in [36], i.e., $|d| < 0.2$ 'negligible', $|d| < 0.5$ 'small', $|d| < 0.8$ 'medium', otherwise 'large'. The results of this test indicated that for almost half of the features the effect size was at least medium. Despite the fact that almost all features are different for each of the stress levels, their effect sizes are small and just a few of them are medium or large for some of the users (details of the statistical results for each feature were omitted due to space constraints). These exploratory results suggest that some of the features (independently of the others) can be used as potential predictors of stress levels. Next, we will use multivariate statistical models and feature selection to find combinations of good discriminative features to detect stress levels. Since we quantified the original stress levels as three different classes {low, medium, high} we will state the problem as a classification problem. Given the set of computed features from the accelerometer data we want to predict the users' self-reported stress levels. In this case we will use multivariate classification models which are discussed in the next section.

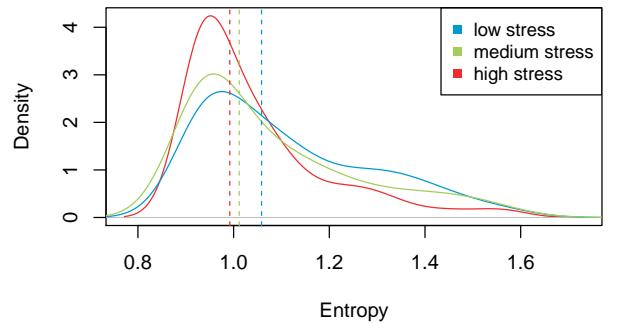


Fig. 2. Estimated density for Entropy feature. Vertical lines represent the median

VI. STATISTICAL MODELS

The results from the exploratory data analysis suggest that we could use some of the features as predictors to classify the different stress levels. For this purpose, we are going to use two classification models namely: 1) Naive Bayes [37] (pp. 90-97) and 2) Decision Trees [38] (ch. 7).

As we discussed earlier, some of the features may increase the performance of the classifiers while others may have the opposite effect. To find good combinations of features to build the models we used a feature selection method called Forward Feature Selection [39] (pp. 207) which consists of adding predictors one by one to the model and at each step the variable that increases performance criteria the most is retained. In this case we used accuracy as the performance criteria.

A. Model Schemes

In recent related works it has been common to build *user-specific* and *general* models to classify stress levels [27], [22]. For the *user-specific* case, individual models are trained and evaluated for each of the users using their own data. The *general model* consists of building the model with data from all the users. This can be done by aggregating all the data from all the users or for each specific user i build a model with the data from all other users j , $j \neq i$ and test the model with the data from user i . The latter approach is sometimes referred to as *leave one person out*. In Lu et al. [22] they also used an hybrid approach called model adaptation which starts with a general model and gets adapted to each individual as more data is available.

Following this methodology, we used the *user-specific* and the *general model* approach. For the *general model* we used the *leave one person out* scheme. We also built *similar-users models* which differ from the general model in that instead of building one model for a user i using observations from all other users $j, j \neq i$, the model for user i is built using observations from just a subset of similar users. The rationale behind this scheme is that for any two users, their behavioural patterns across stress levels may be different. For example, a user may tend to be more active when he is stressed but another one may tend to be more sedentary when stressed.

Building a single model that includes users with different behaviour patterns is not desirable since this will introduce noise. Rather, we may want to build a model for a specific user with data just from similar users. In this case, even if there is not yet enough data to build an *user-specific* model a system could build a model from similar users and start giving feedback until there is sufficient data to build an individual model.

Similar-users Model: Here, the idea is to build a model to predict stress levels for the test user u_t using data from the set of users \mathbb{S} , where \mathbb{S} is the set of users with similar behaviour to u_t . The behaviour of each user will be represented by a single vector \mathbf{b}_i of size $= \binom{|C|}{2} |F|$ where $|C|$ is the number of classes and $|F|$ is the number of features. In this case $\binom{|C|}{2} = 3$ which corresponds to every possible combination of stress levels: *low-medium*, *low-high*, *medium-high*. For each feature we want to know how does the median value changes between the different pairs of stress levels. For example, for one user the difference between $\text{median}(f_{1_{low}}) - \text{median}(f_{1_{high}})$ may be positive but for other user it may be negative where $\text{median}(f_{1_{low}})$ is the median of a specific feature when the stress level is low (and the same applies for all other levels). The behaviour vector \mathbf{b}_i is constructed by computing for each

feature, the difference of the medians between every pair of stress levels.

To find \mathbb{S} we used k-means clustering to group the behaviour vectors \mathbf{b}_i , $i \neq t$ into k groups $G_{1..k}$ and let \mathbb{S} be the group who's centroid has the minimum distance to the behaviour vector of the test user \mathbf{b}_t , i.e., $\mathbb{S} = \arg \min_{G_{1..k}} \text{dist}(\mathbf{b}_t, \text{centroid}(G_i))$. Since u_t is the test user, \mathbf{b}_t is computed using only a random subset $O_{t,p}$ of the total observations of user t where p indicates what percentage of the total observations was taken. The subset $O_{t,p}$ of observations that was used to construct \mathbf{b}_t to find the similar users is discarded when evaluating the model to avoid over-fitting.

The k-means algorithm requires to specify the number k of desired groups. To find a good approximation of k we used the *silhouette* index [40] which is a measure of the quality of the resulting groups. The k-means algorithm is run for $k = 2, 3, \dots, \text{upperbound}$ and the k that maximizes the silhouette index is chosen as the final number of groups. Figure 3 shows an example of the resulting silhouette plot when grouping similar users to build a model for some specific subject when $k = 2$. In this plot each line represents a behaviour vector \mathbf{b} and its length represents its silhouette width $s(i)$. A $s(i)$ close to 1 means that the feature vector i is well clustered, i.e., there is little doubt that i has been assigned to an appropriate group. The overall silhouette index is the average of all $s(i)$ and in this case it was 0.32. It can be seen that some feature vectors had a silhouette width less than or close to 0. This means that it is not clear whether these feature vectors should have been assigned to another cluster. Figure 4 shows the silhouette plot for $k = 3$. In this case the silhouette index was 0.2 which is much lower and in the first cluster almost all data points have a silhouette width close to 0. For $k = 4, 5$ the silhouette index was 0.2 and 0.18 respectively, thus, $k = 2$ was chosen as the number of final clusters for this specific user. Note that the plots have 26 bars (users) instead of the expected 29. This is because some users did not report *high* stress levels and thus they have missing values in their behaviour vectors in which case they were excluded from the clustering phase. On the other hand, if the test user u_t did not report *high* stress levels, the columns with *high* stress levels of the other users' behaviour vectors are truncated and thus, all other 29 users were included in the clustering procedure.

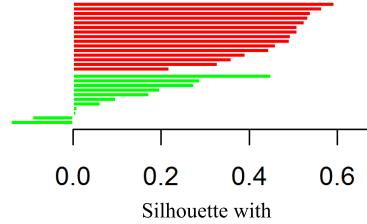


Fig. 3. Silhouette plot for $k = 2$ with resulting silhouette index of 0.32. Line colors represent the different clusters.

B. Ordinal Classification

Typically, classification algorithms assume that the response class is unordered but there are situations in which there is

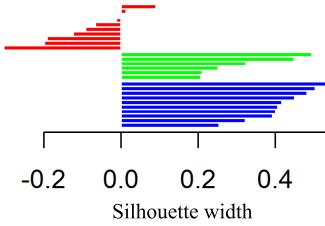


Fig. 4. Silhouette plot for $k = 3$ with resulting silhouette index of 0.2. Line colors represent the different clusters.

a natural ordering of the response variable, i.e., an ordinal class. Ordinal variables are typically found in surveys' responses for example, *Very poor, Poor, Fair, Good, Excellent*. For our case we have: *low < medium < high* stress levels. In order to take into account this ordering information, we also implemented an ordinal classification approach described by Frank & Hall [41] which enables standard classification algorithms to make use of ordering information. This approach consists of transforming a k -class ordinal problem into $k-1$ binary class problems and computing the probability of each of the k ordinal classes. The final prediction is the class with maximum probability. We applied this approach with the Naive Bayes classifier.

From the performance measurement point of view, usually, the classifiers are assessed with measures appropriate for unordered classes. These measures treat all errors as equal, e.g., confusing *low* with *medium* has the same error weight as confusing *low* with *high* but clearly, the latter error should be more severely penalized as discussed in [27]. In [42] several performance measures for ordinal classes were evaluated. For example, *Mean Squared Error* (MSE) is more suitable when the severity of the errors is more important while *Mean Absolute Error* (MAE) is preferred in situations where the tolerance for small errors is lower. Another performance measure is the *Linear Correlation*. A strong correlation between the predictions and the ground truth is an indication of a good classifier. A more optimistic measure is the *Accuracy within n* (ACC1, ACC2,..., ACCn) which allows a wider range of outputs to be considered correct. For example, if the correct output is 4, outputs of 3,4 and 5 are considered as correct for $n = 1$, i.e., ACC1. The usual accuracy measure would be ACC0.

VII. EXPERIMENTS AND RESULTS

In this section we present the results for the three model schemes discussed in Section VI-A: *user-specific*, *general* and *similar-users* models. In the Feature Forward Selection step, for each of the candidate feature subsets 5-fold cross validation is performed in the case of *user-specific* models and *leave one person out* cross validation for the *general* and *similar-users* models. For the *similar-users* model, 50% of the data was used to find the most similar users, i.e., $O_{i,50}$. We used the following performance measures that take into account the ordinal nature of the response variable to evaluate the models: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Pearson

Correlation (Pearson cor), Spearman Correlation (Spearman cor) and Accuracy within 1 (ACC1).

In our experiments, 4 classifiers were used: *Naive Bayes*; *Decision Tree*; *Ordinal Naive Bayes* which uses the approach described in Section VI-B; and as a baseline a *Random* classifier which randomly predicts a class based on their prior probabilities. Table III shows the results for the *user-specific* models. Here we can see that all classifiers (except Random) had a similar overall performance. Note that the ACC1 measure is very optimistic. The Random classifier had an ACC1 = 0.81. this is because an output will be counted as an error only if the prediction is *low* and the actual class is *high* or vice versa. For the *user-specific* case, the 10 most frequently selected features (in descending order) were: Magnitude, Standard Deviation of the 3 axes, Minimum Energy, Maximum of the 3 axes, Peak Magnitude Frequency, Minimum variance Y, Maximum of variance sum, Max Range of the 3 axes, Maximum Mean Energy and Variance sum.

Table IV shows the results for the *general* model scheme. As expected, the overall performance is much lower than that of the *user-specific* scheme. Again all classifiers (except Random) had similar overall performances. The Random classifier had Pearson and Spearman correlations close to 0 while for the other models the correlation was stronger but still weak. The *Ordinal Naive Bayes* classifier did not present any improvement over the traditional *Naive Bayes*. The reason of this lack of improvement when including ordering information may be that in this case the number of classes is just 3 and as suggested by Frank & Hall [41] "ordering information becomes more useful as the number of classes increases".

Table V shows the results for the *similar-users* model for *Naive Bayes* and *Decision tree*. The *Ordinal Naive Bayes* was omitted since it did not present any performance improvement in the previous cases. With respect to the general model, the *similar-users* model had an increase of 8% in accuracy for *Naive Bayes* and 5% for the *Decision Tree*.

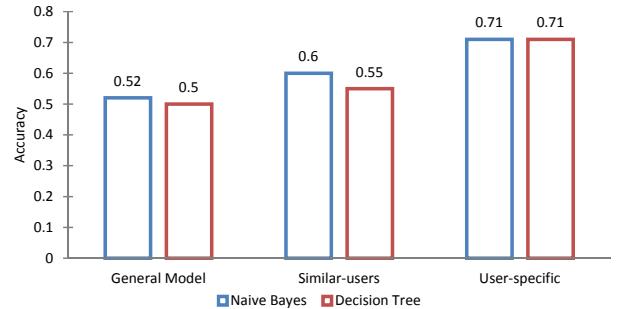


Fig. 5. Comparison between general models, similar users models and user specific models.

VIII. CONCLUSIONS

This work was a first step in evaluating the potential of mobile phones as stress detectors in working environments. The data was collected in an *unconstrained* environment with *unknown* stressors. We used accelerometer data to characterise subjects' behaviour by extracting time domain and frequency

TABLE III
USER-SPECIFIC MODEL RESULTS

	Random			Naive Bayes			Ordinal Naive Bayes			Decision Tree		
	low	medium	high	low	medium	high	low	medium	high	low	medium	high
Sensitivity	0.57	0.41	0.34	0.82	0.65	0.59	0.82	0.62	0.57	0.79	0.66	0.62
Specificity	0.68	0.69	0.8	0.81	0.82	0.91	0.79	0.82	0.91	0.82	0.83	0.9
Precision	0.58	0.41	0.32	0.77	0.65	0.66	0.75	0.65	0.65	0.77	0.67	0.63
Accuracy	0.46			0.71			0.7			0.71		
MAE	0.66			0.33			0.35			0.34		
RMSE	0.96			0.65			0.67			0.68		
Pearson cor	0.24			0.63			0.62			0.61		
Spearman cor	0.25			0.64			0.62			0.62		
ACC1	0.86			0.95			0.95			0.94		

TABLE IV
GENERAL MODEL RESULTS

	Random			Naive Bayes			Ordinal Naive Bayes			Decision Tree		
	low	medium	high	low	medium	high	low	medium	high	low	medium	high
Sensitivity	0.41	0.33	0.24	0.94	0.18	0.2	0.91	0.18	0.22	0.84	0.19	0.28
Specificity	0.58	0.63	0.77	0.3	0.91	0.95	0.33	0.89	0.94	0.39	0.85	0.91
Precision	0.43	0.32	0.23	0.51	0.53	0.55	0.51	0.47	0.54	0.52	0.41	0.47
Accuracy	0.35			0.52			0.51			0.5		
MAE	0.83			0.62			0.61			0.62		
RMSE	1			0.95			0.94			0.94		
Pearson cor	0.01			0.32			0.33			0.31		
Spearman cor	0.01			0.32			0.33			0.31		
ACC1	0.81			0.85			0.86			0.87		

TABLE V
SIMILAR-USERS MODEL RESULTS

	Naive Bayes			Decision Tree		
	low	medium	high	low	medium	high
Sensitivity	0.6	0.58	0.6	0.64	0.64	0.24
Specificity	0.83	0.69	0.86	0.76	0.58	0.95
Precision	0.73	0.5	0.55	0.67	0.44	0.59
Accuracy	0.6			0.55		
MAE	0.45			0.49		
RMSE	0.75			0.76		
Pearson cor	0.52			0.43		
Spearman cor	0.52			0.44		
ACC1	0.94			0.95		

domain features. Then, statistical models were built to classify different self-reported stress levels. For our experiments, we also evaluated an ordinal classification method, which had no improvement in the overall performance, possibly due to the small number of classes (just 3). *User-specific* models performed the best since they are targeted for each specific user but they require more labelled data. On the other hand, *general* models had a lower overall performance but they don't require user specific labelled data which is sometimes tedious and time consuming to record. We proposed a *similar-users* model in which a small amount of labelled data is used to find similar users and a classifier is built. According to our results, this proved to be a middle point between *general* and *user-specific* models, allowing a future system to begin providing feedback to the users on the onset, using *general* model and as more labelled data is available *similar-users* and *user-specific* models could be built. The results we achieved are similar to the results found during literature review, with the difference that in our work we used a single accelerometer sensor only. This could open the possibility to implement a stress recognition system in personal fitness devices, which currently track physical activity only. Our follow-up study will

extend data collection period to several months and include higher number of users in the experiments. Further analysis will focus on analysis of specific situations when the person is handling the phone (such as during phone call, text writing), which may provide a more fine grained insight into the users' behaviour.

ACKNOWLEDGEMENTS

Enrique Garcia-Caja would like to thank Consejo Nacional de Ciencia y Tecnología (CONACYT) and the AAAMi research group at Tecnológico de Monterrey for the financial support in his PhD studies.

REFERENCES

- [1] M. Frese, "The changing nature of work." *An Introduction to Work and Organizational Psychology: a European perspective*. Blackwell Publishers, vol. 1, pp. 424–439, 2000. [Online]. Available: <http://catdir.loc.gov/catdir/toc/fy036/99032109.html>
- [2] "European foundation for the improvement of living and working conditions," *Fourth European Working conditions Survey*, 2005. [Online]. Available: <http://www.eurofound.europa.eu/ewco/surveys/EWCS2005/index.htm>
- [3] "Stress in america." *American Psychological Association*, August 2012. [Online]. Available: <http://apa.org/news/press/releases/stress/index.aspx?tab=2>

- [4] P. Schnall, K. Belkic, P. Landsbergis, and D. Baker, "The workplace and cardiovascular disease," *Occupational Medicine: State of the art - reviews*, no. 15, pp. 24–46, jan 2000.
- [5] P. M. Bongers, C. R. de Winter, M. A. Kompier, and V. H. Hildebrandt, "Psychosocial factors at work and musculoskeletal disease," *Scandinavian Journal of Work: Environment and Health*, no. 19, pp. 297–312, 1993.
- [6] N. Kawakami, T. Tanigawa, A. S., A. Nakata, S. Sakurai, K. Yokoyama, and Y. Morita, "Effects of job strain on helper-inducer (d4+cd29+) and suppressor-inducer (cd4+cd45ra+) tcells in japanese blue-collar workers," *Psychotherapy and Psychosomatics*, no. 66, pp. 192–198, 1997.
- [7] P. A. Thoits, "Self, identity, stress, and mental health." *Handbook of the sociology of mental health*. Springer Netherlands, pp. 357–377, 2013. [Online]. Available: http://link.springer.com/chapter/10.1007/978-94-007-4276-5_18#page-1
- [8] A. Paoli, P. Parent-Thirion, "Working conditions in the acceding and candidate countries." *European Foundation for the Improvement of Living and Working Conditions, Office for Official Publications of the European Communities.*, no. 6, 2003. [Online]. Available: www.eurofound.europa.eu/publications/htmlfiles/ef0306.htm
- [9] D. Hillier, F. Fewell, W. Cann, and V. Shephard, "Wellness at work: Enhancing the quality of our working lives," *International Review of Psychiatry*, vol. 17, no. 5, pp. 419–431, 2005.
- [10] L. He, M. Lech, M. Maddage, and N. Allen, "Stress detection using speech spectrograms and sigma-pi neuron units," in *Natural Computation, 2009. ICNC'09. Fifth International Conference on*, vol. 2, Tianjin, China, 2009, pp. 260–264.
- [11] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 921–926.
- [12] E. Demerouti and A. B. Bakker, "The oldenburg burnout inventory: A good alternative to measure burnout and engagement," *Handbook of stress and burnout in health care*. Hauppauge, NY: Nova Science, 2008.
- [13] R. Kocielnik, M. Pechenizkiy, and S. Natalia, "Stress analytics in education." in *EDM*, 2012. [Online]. Available: http://rkocielnik.com/publications/education_edm2012.pdf
- [14] C. M. Ikebara CS, "Assessing cognitive load with physiological sensors." *Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.
- [15] A. Grunerbl, A. Muaremi, V. Osmani, G. Bahle, S. Ohler, G. Troster, O. Mayora, C. Haring, and P. Lukowicz, "Smartphone-based recognition of states and state changes in bipolar disorder patients." *Biomedical and Health Informatics, IEEE Journal of*, vol. 19, no. 1, pp. 140–148, Jan 2015.
- [16] A. Gruenerbl, V. Osmani, G. Bahle, J. C. Carrasco, S. Oehler, O. Mayora, C. Haring, and P. Lukowicz, "Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients," in *Proceedings of the 5th Augmented Human International Conference on - AH '14*. New York, New York, USA: ACM Press, Mar. 2014, pp. 1–8. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2582051.2582089>
- [17] A. Sano and R. Picard, "Stress recognition using wearable sensors and mobile phones," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, Sep. 2013, pp. 671–676.
- [18] G. Bauer and P. Lukowicz, "Can smartphones detect stress-related changes in the behaviour of individuals?" in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, Mar. 2012, pp. 423–426.
- [19] D. Carneiro, J. C. Castillo, P. Novais, A. Fernández-Caballero, and J. Neves, "Multimodal behavioral analysis for non-invasive stress detection," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13 376–13 389, 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0957417412007816>
- [20] D. Giakoumis, A. Drosou, P. Cipresso, D. Tzovaras, G. Hassapis, A. Gaggioli, and G. Riva, "Using activity-related behavioural features towards more effective automatic stress detection." *Plos One*, vol. 7, no. 9, p. e43571, 2012.
- [21] A. R. Jensen and W. D. R. Jr., "The stroop color-word test: A review," *Acta Psychologica*, vol. 25, no. 0, pp. 36–93, 1966. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/0001691866900047>
- [22] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "StressSense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp '12. New York, NY, USA: ACM, 2012, pp. 351–360. [Online]. Available: <http://doi.acm.org/10.1145/2370216.2370270>
- [23] A. Matic, V. Osmani, A. Maxhuni, and O. Mayora, "Multi-modal mobile sensing of social interactions," in *Pervasive computing technologies for healthcare (PervasiveHealth), 2012 6th international conference on*. IEEE, 2012, pp. 105–114.
- [24] A. Matic, V. Osmani, and O. Mayora, "Trade-offs in monitoring social interactions," *Communications Magazine, IEEE*, vol. 51, no. 7, pp. 114–121, Jul. 2013.
- [25] A. Matic, V. Osmani, and O. Mayora-Ibarra, "Analysis of social interactions through mobile phones," *Mobile Networks and Applications*, vol. 17, no. 6, pp. 808–819, 2012.
- [26] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, "Activity-aware mental stress detection using physiological sensors," in *Mobile Computing, Applications, and Services*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, M. Gris and G. Yang, Eds. Springer Berlin Heidelberg, 2012, vol. 76, pp. 211–230. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-29336-8_12
- [27] A. Muaremi, B. Arnrich, and G. Tröster, "Towards measuring stress with smartphones and wearable devices during workday and sleep," *BioNanoScience*, vol. 3, no. 2, pp. 172–183, 2013. [Online]. Available: <http://dx.doi.org/10.1007/s12668-013-0089-2>
- [28] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. Pentland, "Pervasive stress recognition for sustainable living," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, Mar. 2014, pp. 345–350.
- [29] S. Zhang, P. Murray, R. Zillmer, R. G. Eston, M. Catt, and A. V. Rowlands, "Activity classification using the genea: optimum sampling frequency and number of axes." *Medicine and science in sports and exercise*, vol. 44, no. 11, pp. 2228–2234, 2012.
- [30] Q. Xu, T. L. Nwe, and C. Guan, "Cluster-based analysis for personalized stress evaluation using physiological signals," *Biomedical and Health Informatics, IEEE Journal of*, vol. 19, no. 1, pp. 275–281, Jan 2015.
- [31] "funf open sensing framework. <https://code.google.com/p/funf-open-sensing-framework/source/browse/src/edu/mit/media/funf/probe/builtin/activityprobe.java?r=r55f67cf53180dbba8ed4eddef422737f3abe030e>." accessed January 07 2015.
- [32] S. Mukhopadhyay and G. Ray, "A new interpretation of nonlinear energy operator and its efficacy in spike detection," *Biomedical Engineering, IEEE Transactions on*, vol. 45, no. 2, pp. 180–187, 1998.
- [33] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
- [34] L. Bao and S. Intille, "Activity recognition from user-annotated acceleration data," in *Pervasive Computing*, ser. Lecture Notes in Computer Science, A. Ferscha and F. Mattern, Eds. Springer Berlin Heidelberg, 2004, vol. 3001, pp. 1–17. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-24646-6_1
- [35] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures: Third Edition*. CRC Press, 2003.
- [36] J. Cohen, "A power primer." *Psychological bulletin*, vol. 112, no. 1, pp. 155–159, 1992.
- [37] I. Witten, E. Frank, and M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., ser. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011. [Online]. Available: <http://books.google.com.mx/books?id=5FIEAwyn9aoC>
- [38] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996. [Online]. Available: <http://books.google.it/books?id=2SzT2p8vP1oC>
- [39] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, ser. Springer Texts in Statistics. Springer New York, 2014. [Online]. Available: <http://books.google.it/books?id=at1bmAEACAAJ>
- [40] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.
- [41] E. Frank and M. Hall, "A simple approach to ordinal classification," in *Machine Learning: ECML 2001*, ser. Lecture Notes in Computer Science, L. De Raedt and P. Flach, Eds. Springer Berlin Heidelberg, 2001, vol. 2167, pp. 145–156. [Online]. Available: http://dx.doi.org/10.1007/3-540-44795-4_13
- [42] L. Gaudette and N. Japkowicz, "Evaluation methods for ordinal classification," in *Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science, Y. Gao and N. Japkowicz, Eds. Springer Berlin Heidelberg, 2009, vol. 5549, pp. 207–210. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-01818-3_25